Sadan/Sadri Parallel Text Corpus:
Linguistic Features and Structures

SADAN/SADRI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES

Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition.

Authors:
Ankita Tiwari
Dr. Rejitha K. S
Dr. Narayan Choudhary
Prof. Shailendra Mohan

Linguistic Data Consortium for Indian Languages
Central Institute of Indian Language
Mysuru, India-570006

भारतीय भाषा संस्थान
उच्चतर शिक्षा विभाग, शिक्षा मंत्रालय, भारत सरकार, मानसगंगोत्री, हुन्सूर रोड, मैसूरु-570006, कर्नाटक
CENTRAL INSTITUTE OF INDIAN LANGUAGES
Department of Higher Education, Ministry of Education, Government of India
Manasagangotri, Hunsur Road, Mysuru-570006, Karnataka
www.ciil.gov.in :: e-✉ ciil.publication@gmail.com :: ☎ +91-821-234-5182

Sadan/Sadri Parallel Text Corpus: Linguistic Features and Structures

Authors: Ankita Tiwari, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Radheshyam Lakra and Bibha Rani Topno

e-ISBN: 978-81-69175-28-9
CIIL Publication No.: 1738

First published: AD 2026 March :: Phalguna 1947 Śaka

Publisher: Prof. Shailendra Mohan, Director, CIIL

Printing & Publication Unit
Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit
B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit
M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover design: Ankita Tiwari

TABLE OF CONTENTS

SADAN/SADRI PARALLEL TEXT CORPUS

Ankita Tiwari

1.1 LDC-IL PARALLEL TEXT CORPUS

LDC-IL has developed parallel corpus in Indian mother tongues. India has 270 mother tongues as per 2011 census. Most of these mother tongues have no written text readily available and they are mostly oral languages. Most of them do not have any script of their own or use one or more than one dominant script used in that region. Following the requirements of the NEP-2020, LDC-IL initiated to collect sentences from different mother tongues that could give an idea about the basic structure of these 270 mother tongues.

A parallel corpus is a collection of texts in two or more languages that are carefully aligned sentence wise in pairs and placed either side by side or matched with a common sentence/phrase identification number. It is widely used in linguistics, translation studies, and Natural Language Processing. Researchers use it to compare vocabulary, grammar, and meaning across languages, while computational systems rely on it to train statistical and neural machine translation models. Parallel corpora also support language learning by providing real translation examples and enable cross-language information retrieval, allowing users to access information in multiple languages efficiently.

LDC-IL developed a framework to accommodate all the linguistic features of mother tongues. The sentences were collected from Indian scheduled languages especially from descriptive grammar books. The sentences cover the general grammatical and structural types like adjectival phrases, adverbial phrases, compounding, Interrogative sentences, Imperative sentences, different speech types, statements, structural question, coordination, subordination and so on. Moreover, it includes linguistic features such as use of adposition, anaphora, derivational morphology, inflectional morphology and language specific features like emphasis, equative, possession, reciprocation, reflexive expression and so on. For more details on the source collection for this parallel corpus, please refer to the LDC-IL Corpus Insights (2025).

This parallel corpus can be used in AI development by enabling the creation of Indic multilingual foundation models, develop cross-lingual knowledge graphs and create low-resource language adapters. By providing high-quality linguistic data, it strengthens natural language understanding and generation. As India rapidly advances in the digital age, it is essential to preserve and promote undocumented mother tongues by bringing them onto digital platforms, ensuring they are sustained and able to thrive for future generations.

Low-resource languages have a significant impact on both AI development and digital world. When AI systems are trained mostly on high-resource languages, they become biased toward those linguistic and cultural contexts. Including low-resource languages helps build stronger multilingual foundational models. Exposure to diverse grammar structures, vocabulary, and linguistic patterns improves cross-lingual learning and generalization. Language carries cultural knowledge. Incorporating low-resource languages enriches AI systems with diverse worldviews, idioms, and traditional knowledge systems. If a language is not supported digitally, its speakers may be excluded from online education, e-governance, healthcare services, and digital commerce. When digital tools support local languages, more people can participate in the economy. Local languages often contain indigenous knowledge about agriculture, medicine, environment, and culture and therefore bringing them into digital ecosystem safeguards the linguistic treasures that language might contain.

Supporting low-resource languages makes AI more inclusive, intelligent, and representative of linguistic diversity. In India, where hundreds of mother tongues co-exist, investing in low-resource language technologies is essential for equitable digital growth.

1.2 CHALLENGES IN CREATING A PARALLEL CORPUS

Creating a Parallel Corpus for mother tongues in India involves several linguistic, technical, and socio-cultural challenges. The major difficulty is the lack of standardized orthography in many regional and tribal mother tongues. Most of the mother tongues use the prominent script used in their region. Some mother tongues may use multiple scripts or even a newly created writing systems, which complicates the corpus development. Another challenge is the scarcity of written and digitized resources. Many mother tongues are used for daily communication only.

Dialectal variation is another serious issue, as a single mother tongue may have numerous regional varieties with differences in vocabulary, pronunciation, and grammar. This makes it difficult to ensure consistency and representativeness in the corpus. Accurate translation is a concerned matter. Proficient bilingual speakers may be limited for low-resource languages. Maintaining semantic equivalence across mother tongues with different cultural contexts can be challenging. Sometimes equivalent words are not available due to the socio-cultural differences. These factors affect large-scale parallel corpus development in Indian mother tongues. The creation of reliable and balanced parallel corpora for Indian mother tongues is a complex and resource-intensive task.

1.3 SADAN/SADRI: A BRIEF INTRODUCTION

Sadan/Sadri as (ISO 639-3: sck),[1] is an Indo-Aryan language mainly spoken in the Chhota Nagpur Plateau region. It is also spoken in several other states, including Bihar, Chhattisgarh, West Bengal, Odisha and Assam. The 2011 Census of India reports a total of 43,45,677 speakers of Sadan/Sadri in India.[2] Sadan/Sadri is called by many different names according to the regions, as it is listed by Simons et al. (2017) Chhota Nagpuri, Dikku Kaji, Ganwari, Gauuari, Gawari, Goari, Jharkhandhi, Nagpuri, Nagpuria, Sadana, Sadani, Sadari, Sadati, Sadhan, Sadhari, Sadna, Sadrik, Santri, Siddri, Sradri.

Grierson (1903: 42) mentioned that the "Bhojpuri has three main varieties – the standard, the western and Nagpuri". In the western studies Sadn/Sadri is considered as the dialect of Bhojpuri (Peterson (n.d.)). Rev. Bukaut (1906) has different view in this regard; by saying that Sadan/Sadri must be considered as a fourth form of Bihari (Keshari 2012: 218). It was Grierson who discussed about the Nagpuri as the eastern Magahi. Due to which the census report of 1961 classified Nagpuri together with the eastern Magahi (Grierson 1903). Finally, the Census of India report 2001 has put Sadan/Sadri under the dialect of Hindi.

[1] https://iso639-3.sil.org/code/sck
[2] Albin Rico Xalxo, Descriptive Analysis of Tense and Aspect in Sadri

1.4 SADAN/SADRI: A BRIEF LINGUISTIC OVERVIEW

Sadan/Sadri follows the typical Subject–Object–Verb (SOV) word order found in most Indo-Aryan languages. In simple declarative sentences, the subject usually appears at the beginning, the object follows– often marked by case markers such as the accusative *ke*– and the verb is placed at the end. The verb agrees with the subject in person and number. Although SOV is the standard structure, the language allows some flexibility in word order because grammatical relations are clearly marked through case endings and verbal inflections. As a result, alternative patterns such as OSV or VSO may occur without affecting meaning. In many cases, subject pronouns may be omitted (pro-drop) because verbal inflections provide sufficient information about the subject's person and number. The absence of grammatical gender marking in the language contributes to its relatively simpler morphological structure compared to other Indo-Aryan languages.

Sadan/Sadri shows clear evidence of language contact in its vocabulary. As a regional lingua franca in the Chota Nagpur Plateau, it has incorporated words from Austroasiatic Munda languages as well as from nearby Indo-Aryan languages, including Hindi, Maithili, Bhojpuri and Bengali. At present, Devanagari serves as the primary script used for writing Sadri/Sadan.[3]

1.5 SUMMARY OF SADAN/SADRI PARALLEL CORPUS

The Sadan/Sadri Parallel Text Corpus has been developed using Hindi as the source language. This Corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, Apatani, Are, Assamese, Awadhi, Bagri Rajasthani, Bagri, Bagheli/Baghel Khandi, Balti, Bengali, Bhadrawahi, Bharmauri/Gaddi, Bhojpuri, Bilaspuri Kahluri, Bodo, Brajbhasha, Bundeli/Bundel Khandi, Chakru/Chokri, Chambeali/Chamrali, Chang, Chhattisgarhi, Chirr, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, Gujari, Halabi, Handuri, Hara/Harauti, Haryanvi, Hindi Multani, Hindi, Irula/Irular Mozhi, Kabui, Kachchhi, Kangri, Kannada, Karbi/Mikir, Kashmiri, Khandeshi, Khari Boli, Khasi, Khezha, Khiemnungan, Khortha/Khotta, Kisan, Kodava, Kokbarak, Kolami, Koli, Kom, Konda, Konkani, Konyak, Koya, Kudubi/Kudumbi, Kuki, Kurmali Thar, Ladakhi, Lepcha, Liangmei, Limbu, Lotha, Lyngngam,

[3] https://grokipedia.com/page/Nagpuri_language

Magadhi/Magahi, Maithili, Malayalam, Malvi, Manipuri, Mao, Mara, Maram, Marathi, Maring, Mech/Mechhia, Mewari, Mewati, Miri/Mising, Mishmi, Mizo, Mongsen, Monpa, Mundari, Muwasi, Nawait, Nepali, Nimadi, Nissi, Nocte, Odia, Pahari, Paite, Palmuha, Pania, Pawari/Powari, Phom, Pnar/Synteng, Pochury, Poula, Punjabi, Purkhi, Rai, Rajasthani, Reang, Rengma, Rongmei, Sambalpuri, Sangtam, Sanskrit, Santali, Saurashtra/Saurashtri, Sema, Shina, Sindhi, Sirmauri, Sugali, Surjapuri, Talgalo, Tamil, Tangkhul, Telugu, Thado, Tibetan, Tikhir, Tripuri, Tulu, Urdu, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

The Sadan/Sadri dataset comprises 5,332 sentences/phrases, systematically structured according to 159 grammatical categories. The Sadan/Sadri portion of the corpus contains 31,676 words and a total of 1,44,910 characters. Moreover, the corpus is aligned with 146 Indian mother tongues in addition to English. Altogether, the multilingual parallel corpus encompasses 4,404,845 words (over 4.4 million tokens) and 23,374,289 characters (approximately 23.3 million).

1.6 REFERENCE

[1] Albin Rico Xalxo. (June, 2018). A Descriptive Analysis of Tense and Aspect in Sadri. Language in India (Vol 18). 1930-2940: https://www.languageinindia.com/june2018/albintenseaspectsadri1.pdf

[2] https://censusindia.gov.in/nada/index.php/catalog/42561

[3] https://iso639-3.sil.org/code/sck

[4] https://grokipedia.com/page/Nagpuri_language

[5] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.